# Accurate spectral envelope estimation for articulation-to-speech synthesis

# Accurate spectral envelope estimation for articulation-to-speech synthesis

*Yoshinori Shiga, Simon King*

Centre for Speech Technology Research

University of Edinburgh

# Overview

1. Articulatory-acoustic mapping
2. Problem in parameterization
3. Harmonic-based approach
4. Piecewise linear approximation
5. Advantage of the proposed approach
6. Discussion
7. Conclusion

# Articulatory-acoustic mapping

- Research objective
  - Modifying the acoustic characteristics of speech in articulatorily-meaningful ways
  - Maintaining aspects of the signal relating to speaker identity, and with high signal quality

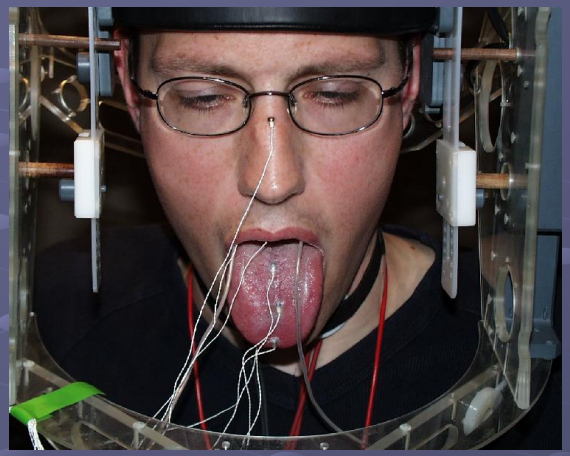
    ➡ Contributing toward synthesising speech in various speaking styles, and different dialects and languages

# Articulatory-acoustic mapping

- Points dealt with in this talk:
  1. A mapping of articulation to the vocal tract transfer function (VTTF) using the actual measurement of articulators
  2. Accurate VTTF estimation based on the articulatory data for high quality speech synthesis

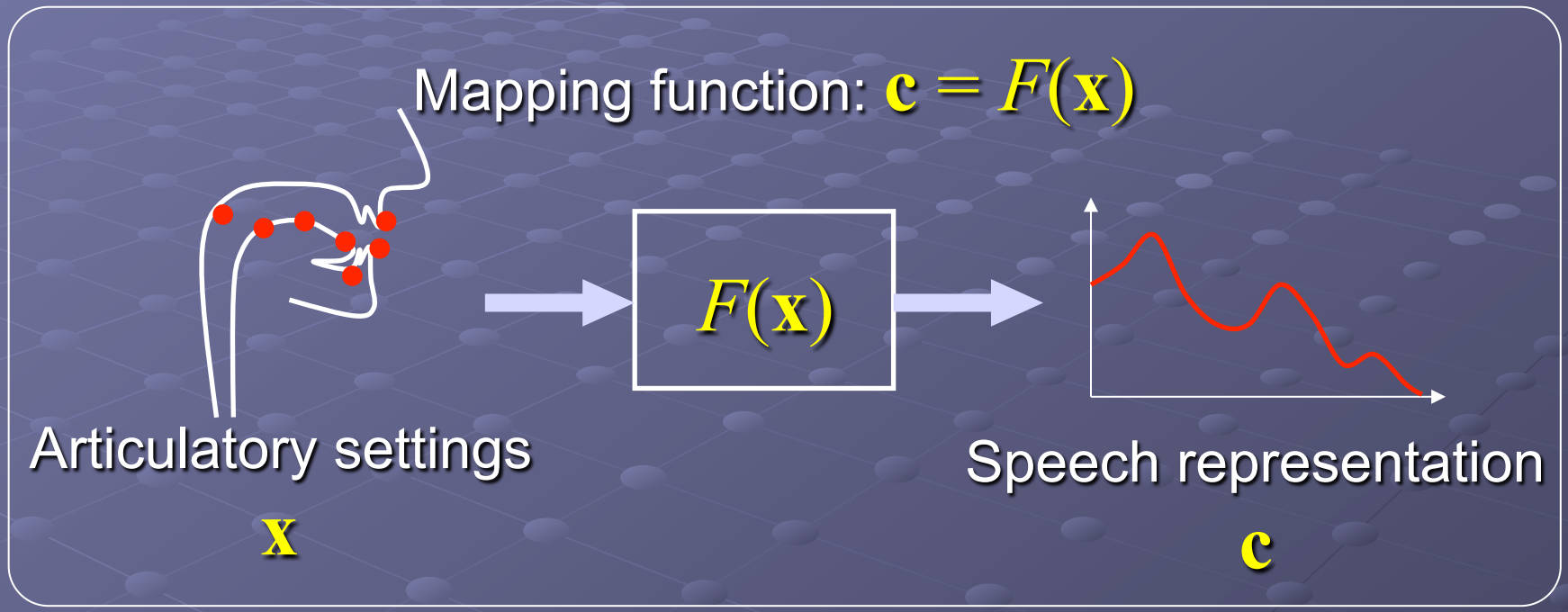
# Articulatory-acoustic mapping

- The articulator positions are measured using the electro-magnetic articulograph (EMA) system.



- EMA data were first applied to articulatory-acoustic mapping by *Kaburagi et al.* (1998)
  - Based on the search of a database composed of pairs of articulator positions and speech spectra.

# Articulatory-acoustic mapping



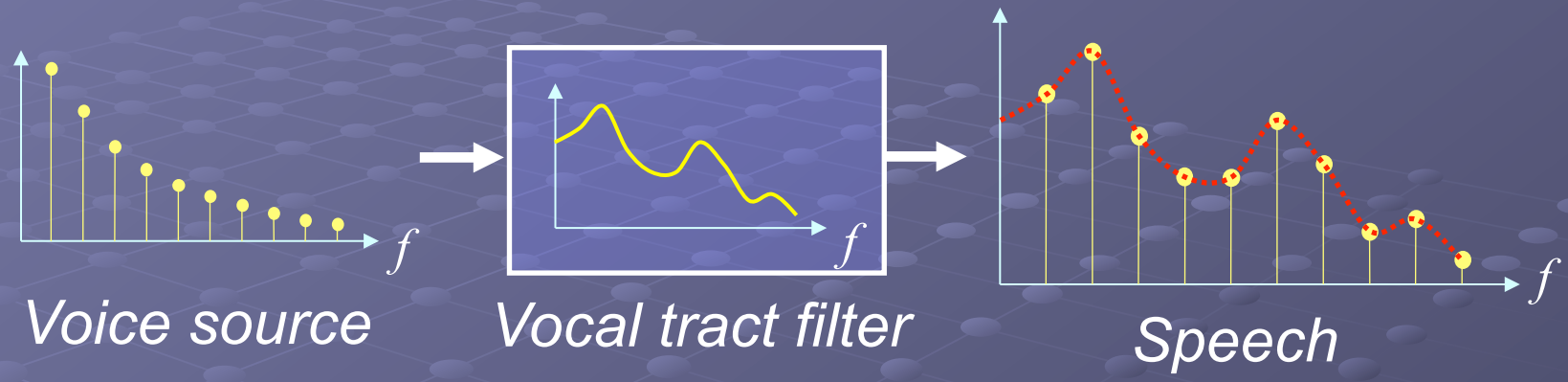

For $F(\mathbf{x})$

- Codebook (VQ)
- Gaussian mixture model
- Neural network, etc

# Problem in parameterization

- Conventional parameter-based synthesis still has many perceptible artefacts.
  - Just applying a common parameterisation to the mapping further degrades speech quality.
- A new parameterisation framework is necessary.

# Problem in parameterization



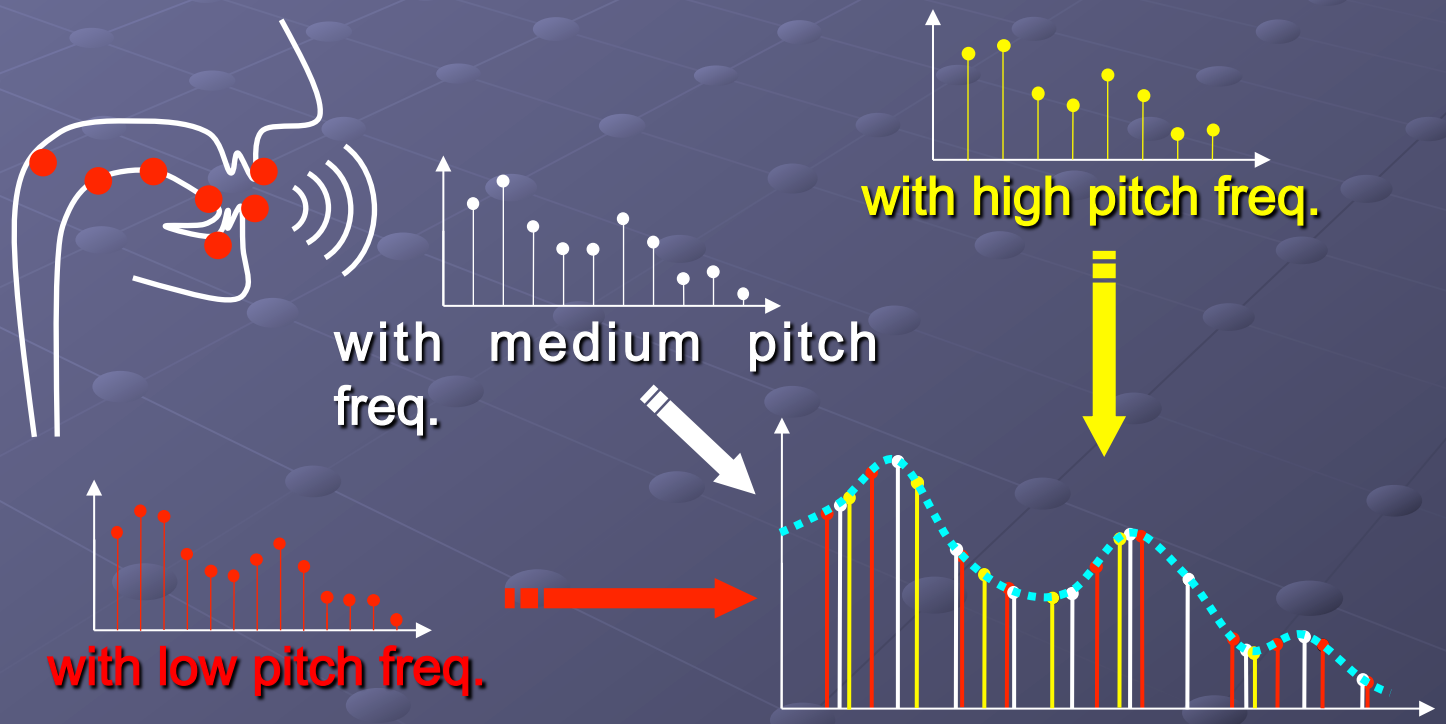

- Envelopes are interpolated into sections between adjacent harmonics.
  - The spacing of harmonics restricts the frequency resolution of spectral envelopes.
  - Harmonic peaks reflect the real VTTF, but the interpolated sections do not.

# Problem in parameterization

- The VTTF should be more complicated because of the intricate vocal tract shape.
  - So, speech generated by parameter-based synthesis is intelligible enough, but unnatural.

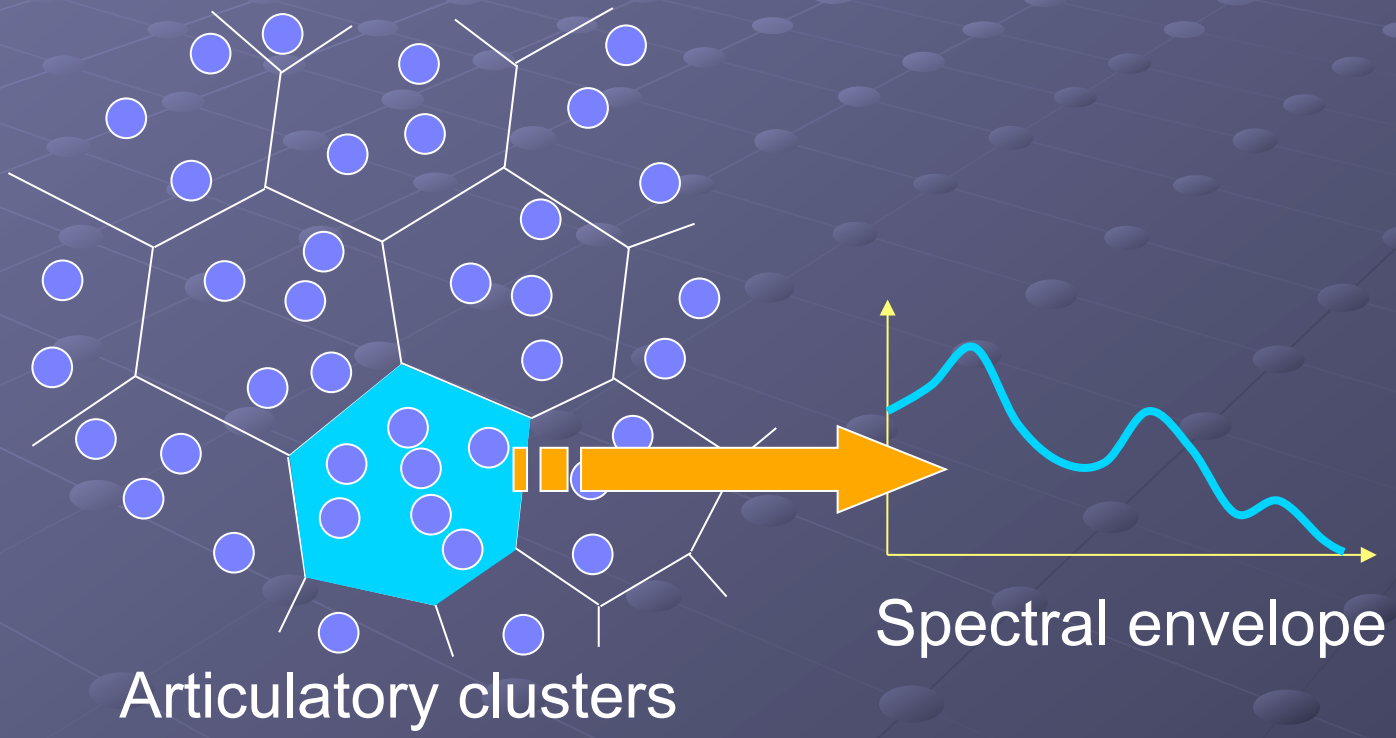
# Harmonic-based approach

- *Solution*: estimating a spectral envelope by collecting sets of harmonics produced using similar articulatory configurations



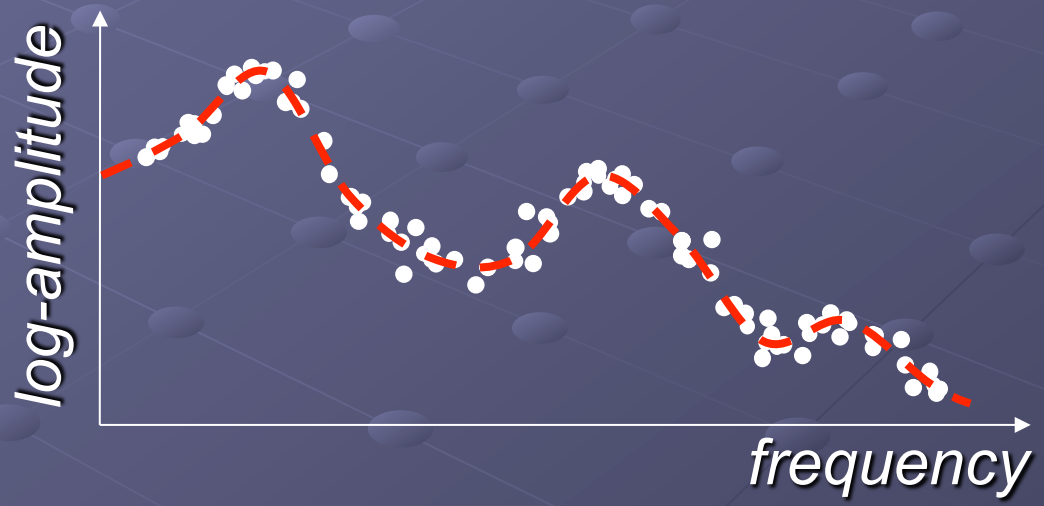
# Harmonic-based approach

- LGB clustering in the articulatory space



# Harmonic-based approach

- Forming a spectral envelope from all the harmonics of all the frames in each cluster
  - An extension of cepstral smoothing based on the least square method (*Galas et al.*, 1990)



# Harmonic-based approach

- Finding cepstrum $\boldsymbol{c}_a^{(i)}$ and $\boldsymbol{c}_p^{(i)}$ representing the power envelope and phase envelope of cluster $C^i$

*For power spectral envelope*

$$E_a^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)} \right]^{\mathrm{T}} \mathbf{W}_k \left[ \boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)} \right]$$

*For phase spectral envelope*

$$E_p^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)} \right]^{\mathrm{T}} \mathbf{W}_k \left[ \boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)} \right]$$

# Harmonic-based approach

- Experiment
  - MOCHA-TIMIT corpus (*Wrench* 2001)
    - Female speaker (fsew0), 460 TIMIT sentences
    - Sampling rate: 16 kHz for speech, 500 Hz for EMA
  - Laryngograph signals for extracting voiced sections.
  - Harmonics estimation: weighted least squares method (*Stylianou,* 2001)
  - 20 ms Hanning window, 8 ms frame spacing
  - 78876 frames for training, 8332 frames for testing

# Harmonic-based approach

- Evaluation
  - For evaluating accuracy only at harmonic frequencies, we introduced *harmonic power distortion* $D_a$, and *harmonic phase distortion* $D_p$.

$$D_a[\mathrm{dB}] = \frac{20}{\ln 10}\sqrt{\frac{1}{M}\sum_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(R(k))})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(R(k))})}$$

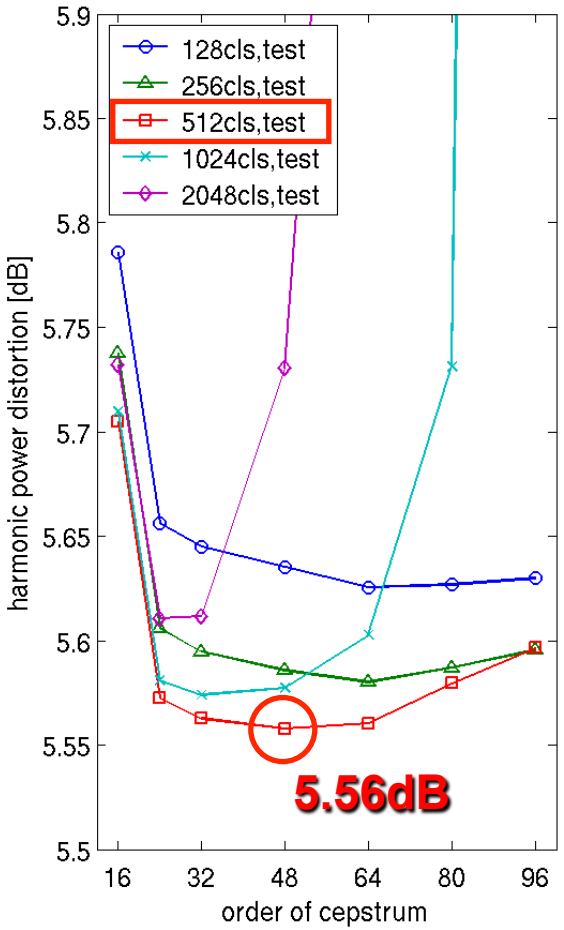
$$D_p[\mathrm{rad}] = \sqrt{\frac{1}{M}\sum_k (\boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(R(k))})^T \mathbf{W}_k (\boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(R(k))})}$$

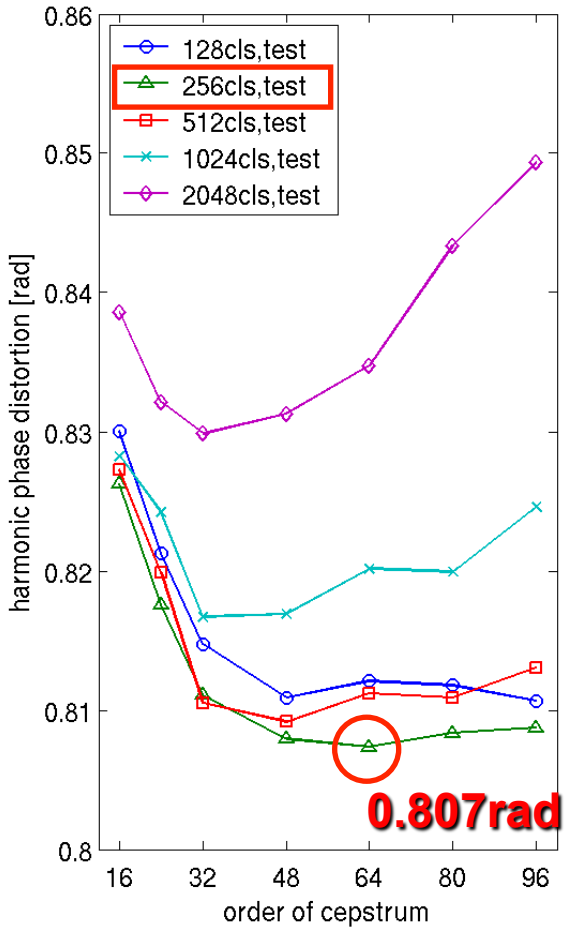
$$\text{frame } k \in C^i \Leftrightarrow i = R(k)$$

$M$ : number of frames evaluated

# Harmonic-based approach

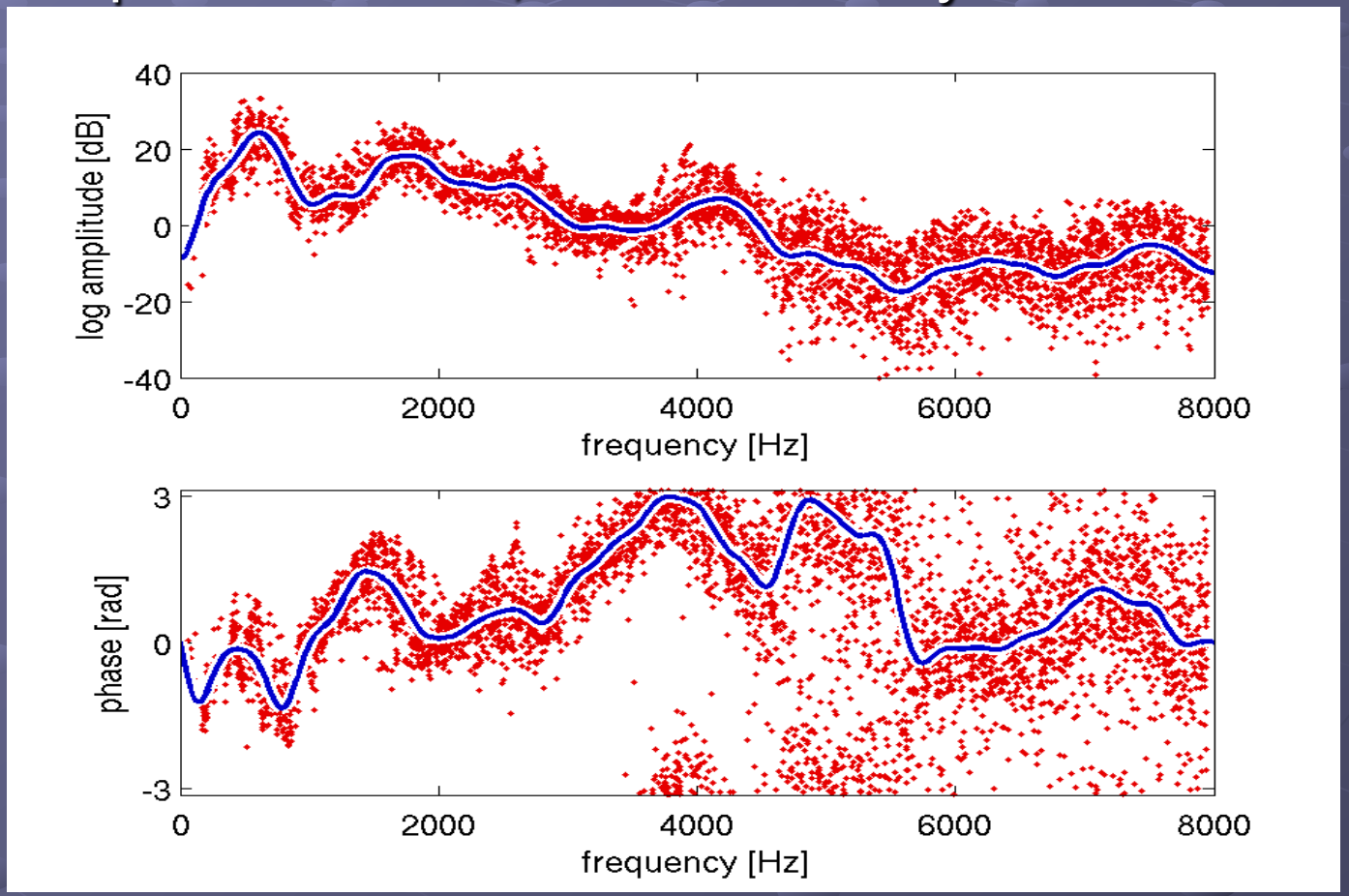
- Experimental results: *harmonic distortions*



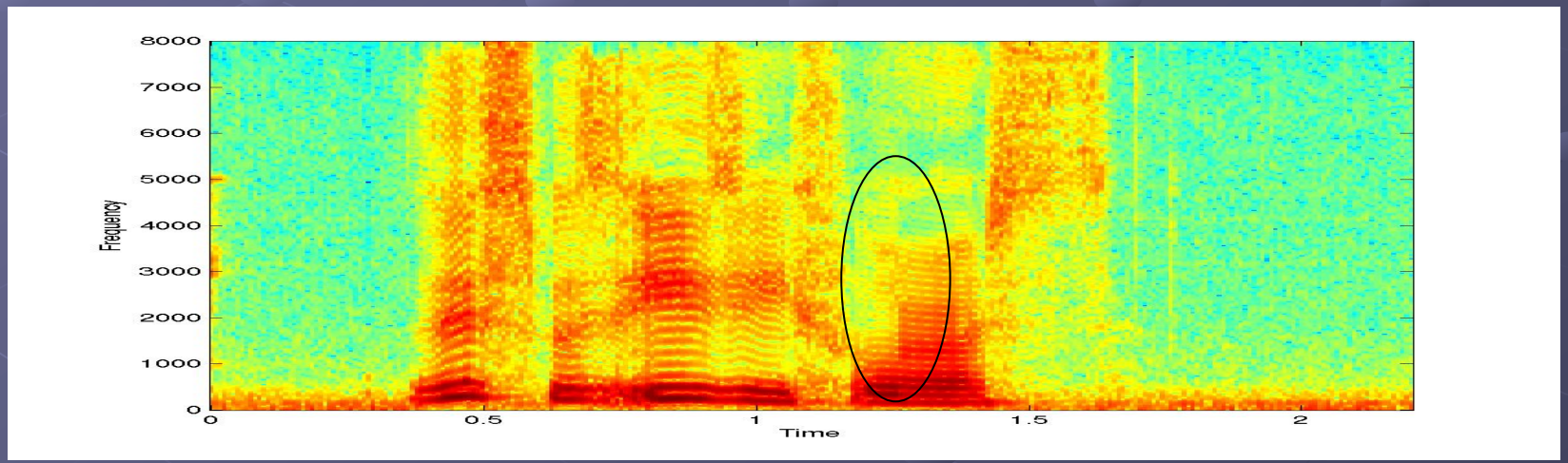
# Harmonic-based approach

- Experimental results: *estimated spectra*
  - Cepstral order 48, 512 articulatory clusters



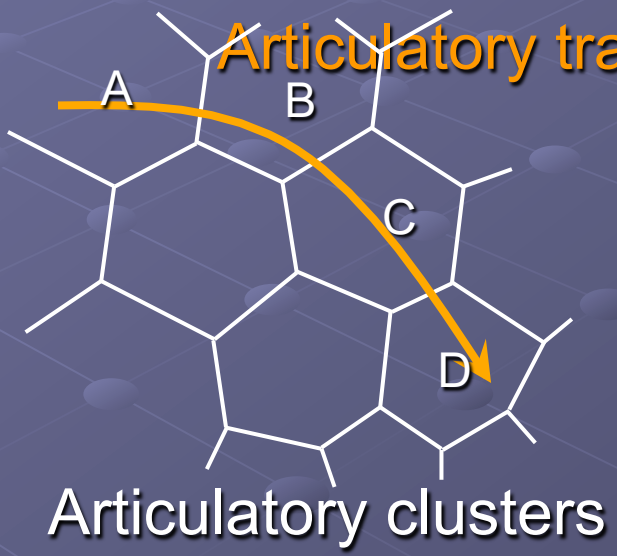
# Harmonic-based approach

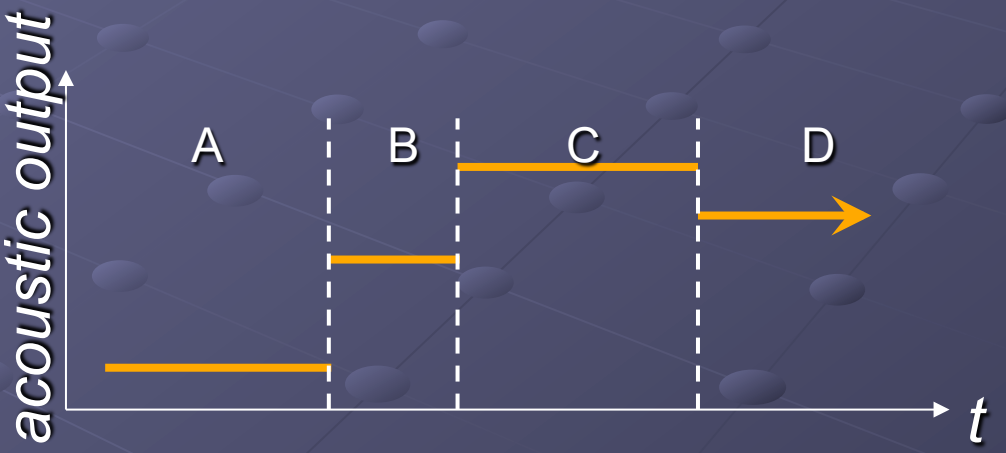
- More detailed envelopes represented using higher order cepstra
- However, articulation is not completely the same in all points within a single cluster.
  - Sufficient accuracy cannot be obtained.
  - VT response changes discontinuously across the cluster boundaries.



# Harmonic-based approach

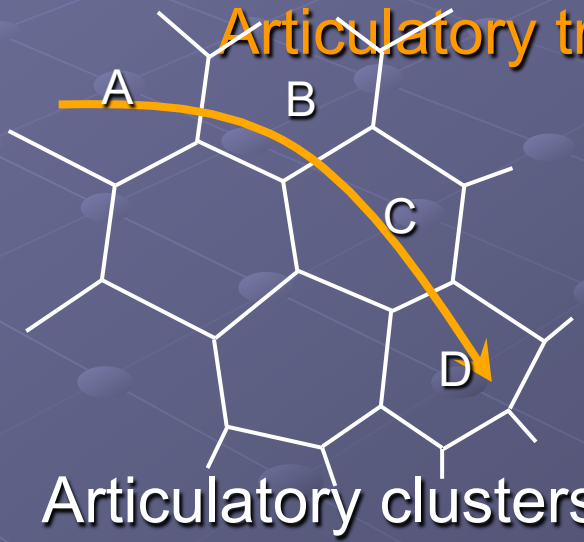
- Our first assumption
  - All articulatory settings in the same cluster correspond to the same spectral envelope.

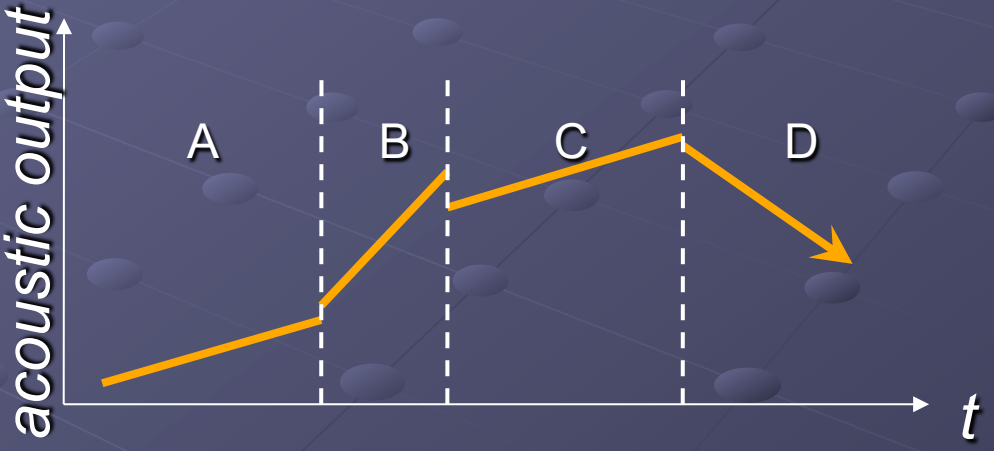


# Harmonic-based approach

- New assumption
  - For each cluster there is a linear relationship between articulatory and acoustic parameters.



# Piecewise linear approximation

- Piece-wise linear assumption

$$\boldsymbol{c}_a^{(i)} = \boldsymbol{q}^{(i)} + \mathbf{U}^{(i)} \boldsymbol{x}_k, \quad \boldsymbol{c}_p^{(i)} = \boldsymbol{r}^{(i)} + \mathbf{V}^{(i)} \boldsymbol{x}_k$$

$\boldsymbol{x}_k$ : articulatory vector of frame $k$

$\boldsymbol{q}$, $\mathbf{U}$, $\boldsymbol{r}$, $\mathbf{V}$: linear regression coefficients

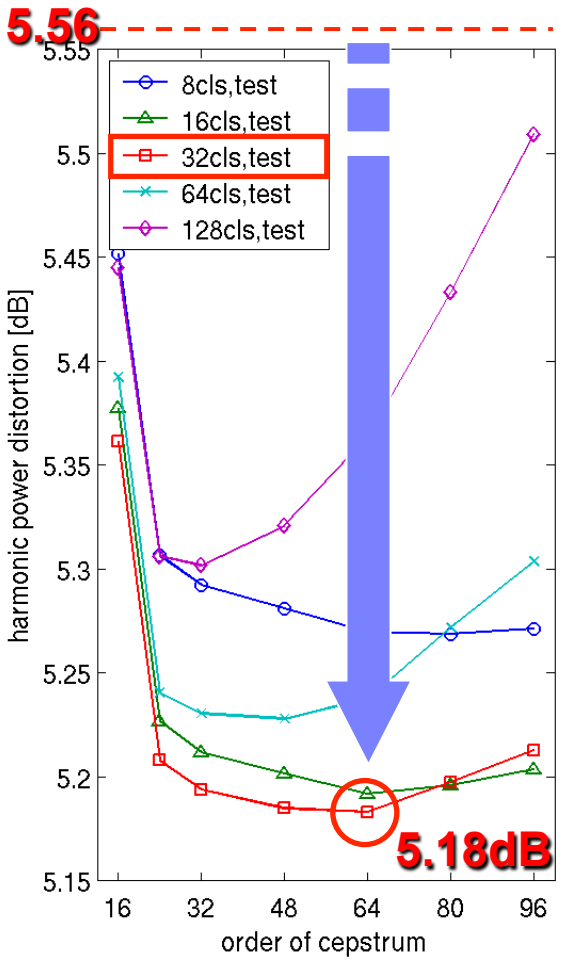
Rewriting the formulae,

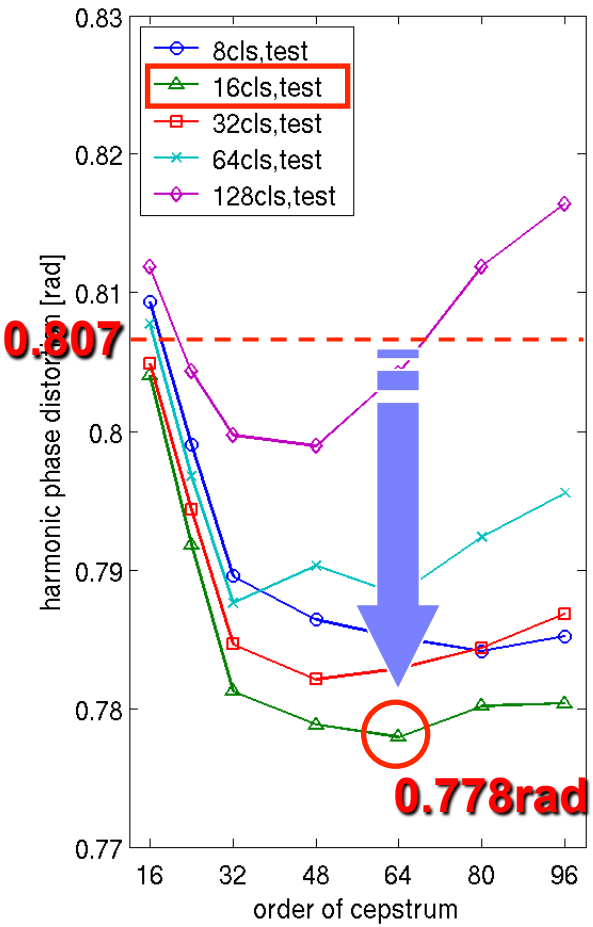
$$E_a^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{y}_k - \boldsymbol{\Gamma}_k \boldsymbol{u}_k^{(i)} \right]^{\mathrm{T}} \mathbf{W}_k \left[ \boldsymbol{y}_k - \boldsymbol{\Gamma}_k \boldsymbol{u}_k^{(i)} \right]$$

$$E_p^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{\theta}_k - \boldsymbol{\Delta}_k \boldsymbol{v}_k^{(i)} \right]^{\mathrm{T}} \mathbf{W}_k \left[ \boldsymbol{\theta}_k - \boldsymbol{\Delta}_k \boldsymbol{v}_k^{(i)} \right]$$

# Piecewise linear approximation

- Experimental results: *harmonic distortions*



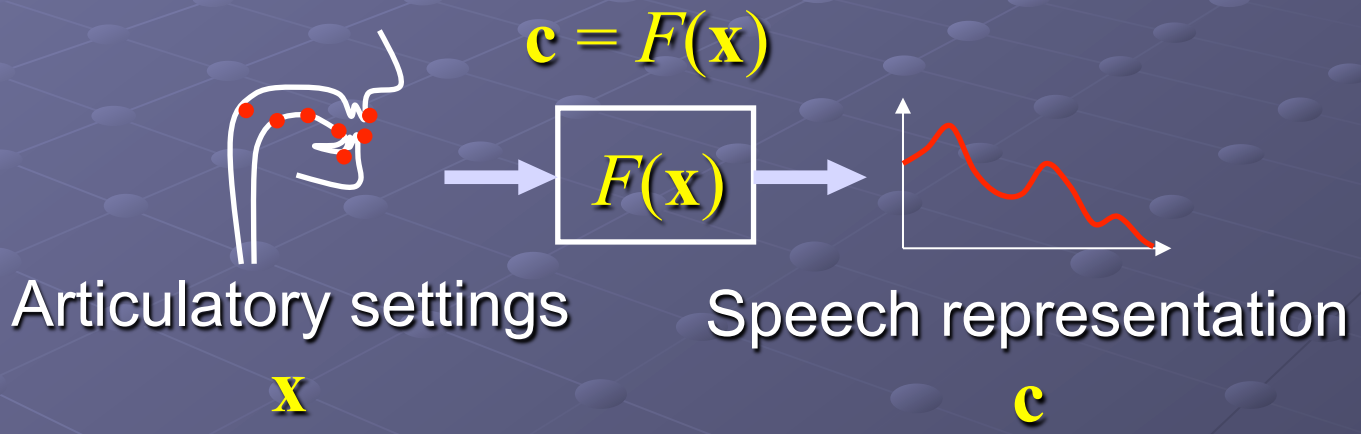



# Summary (1)

- The distortions are minimized when the order is 48-64 (3.0-4.0ms in quefrency).
  - The proposed method can represent more detailed envelopes using high-quefrency cepstral elements.
- The piecewise linear method is more accurate and requires a smaller number of clusters than our first method.
  - The piecewise linear method is more suitable for our proposed parameterization.

# Advantage of the proposed harmonic-based approach

- Obviously, these correspond to piecewise constant and piecewise linear mapping.




- For comparison, we applied a commonly used cepstral-domain criterion as a conventional method.

# Advantage of the proposed harmonic-based approach

- Comparison with a conventional method
  - Cepstral-domain criteria for power

*observed cepstra*

$$E_{PC}^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{c}_k - \boldsymbol{c}_a^{(i)} \right]^{\mathrm{T}} \left[ \boldsymbol{c}_k - \boldsymbol{c}_a^{(i)} \right]$$

$$E_{PL}^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)} \boldsymbol{x}_k) \right]^{\mathrm{T}} \left[ \boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)} \boldsymbol{x}_k) \right]$$

  - Minimum phase was computed from $\boldsymbol{c}_a^{(i)}$ .

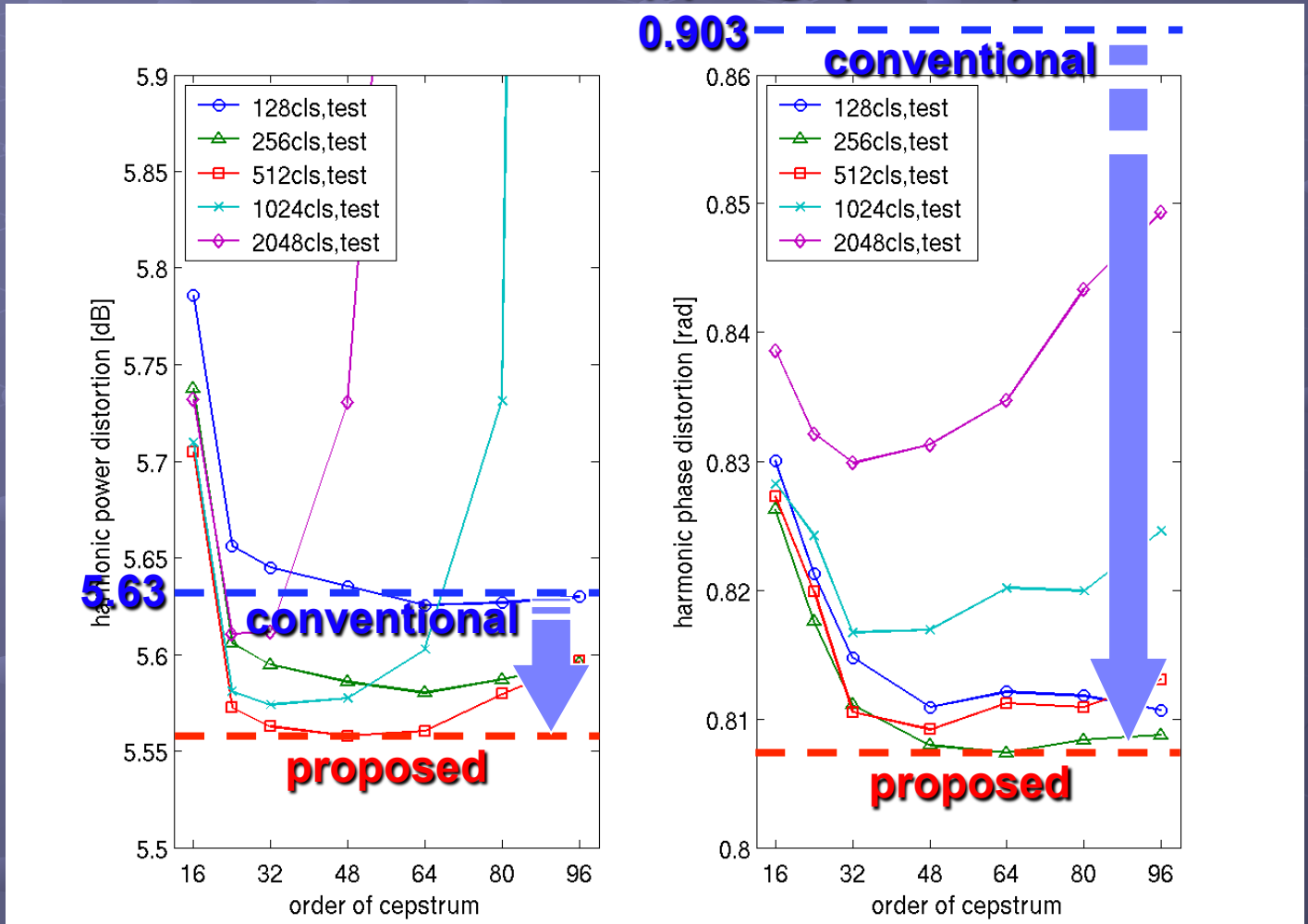
# Advantage of the proposed harmonic-based approach

- Comparison in the harmonic distortion

| | *proposed* | *conventional* |
|---|---|---|
| *Piecewise constant mapping* (PCM) | 5.56 dB (order 48 / 512 cls) | 5.63 dB (order 20 / 512 cls) |
| | 0.807 rad (order 64 / 256 cls) | 0.903 rad (order 20 / 512 cls) |
| *Piecewise linear mapping* (PLM) | 5.18 dB (order 64 / 16 cls) | 5.27 dB (order 20 / 32 cls) |
| | 0.778 rad (order 64 / 32 cls) | 0.887 rad (order 20 / 32 cls) |

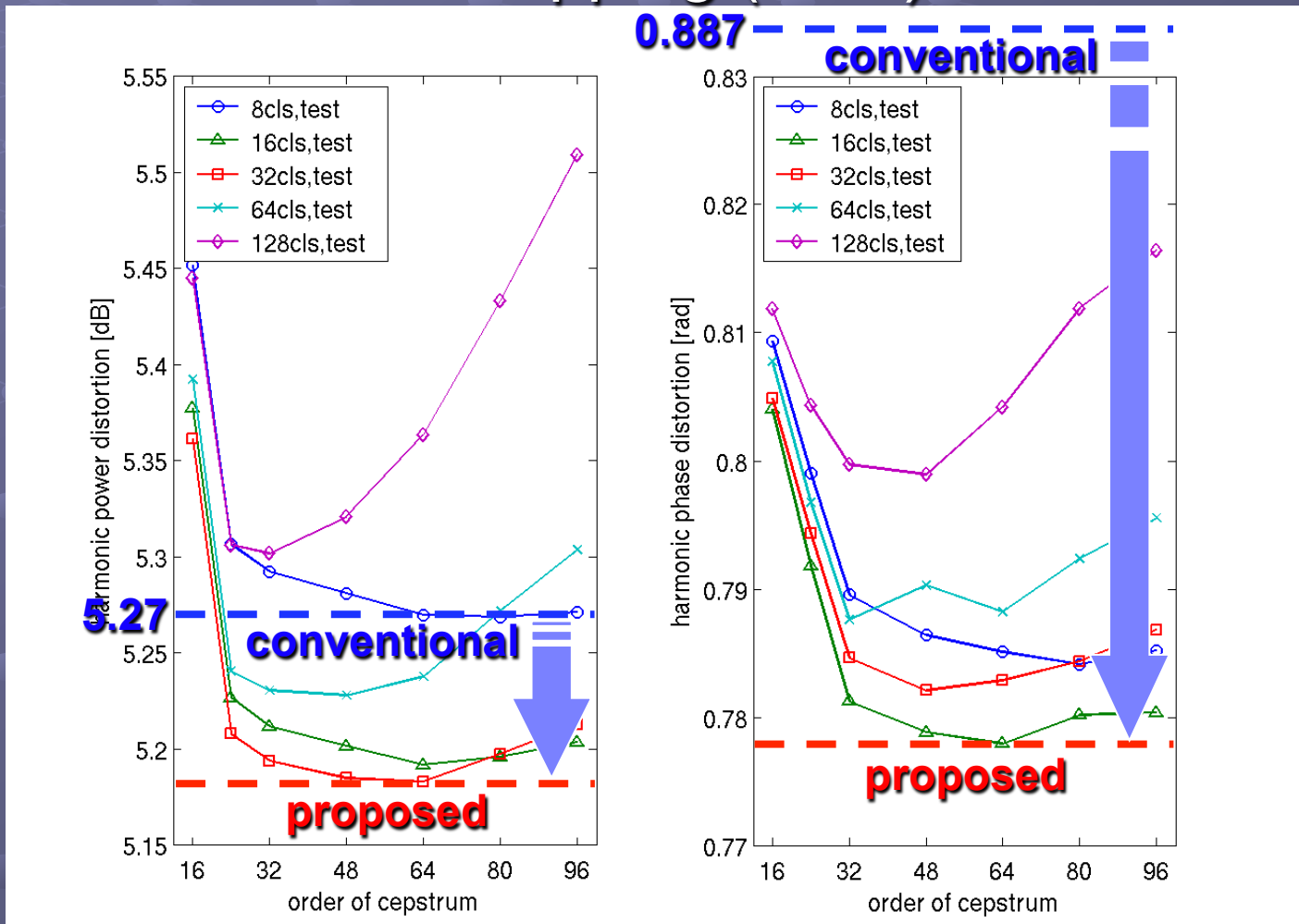
# Advantage of the proposed harmonic-based approach

- Piecewise constant mapping (PCM)



# Advantage of the proposed harmonic-based approach

- Piecewise linear mapping (PLM)

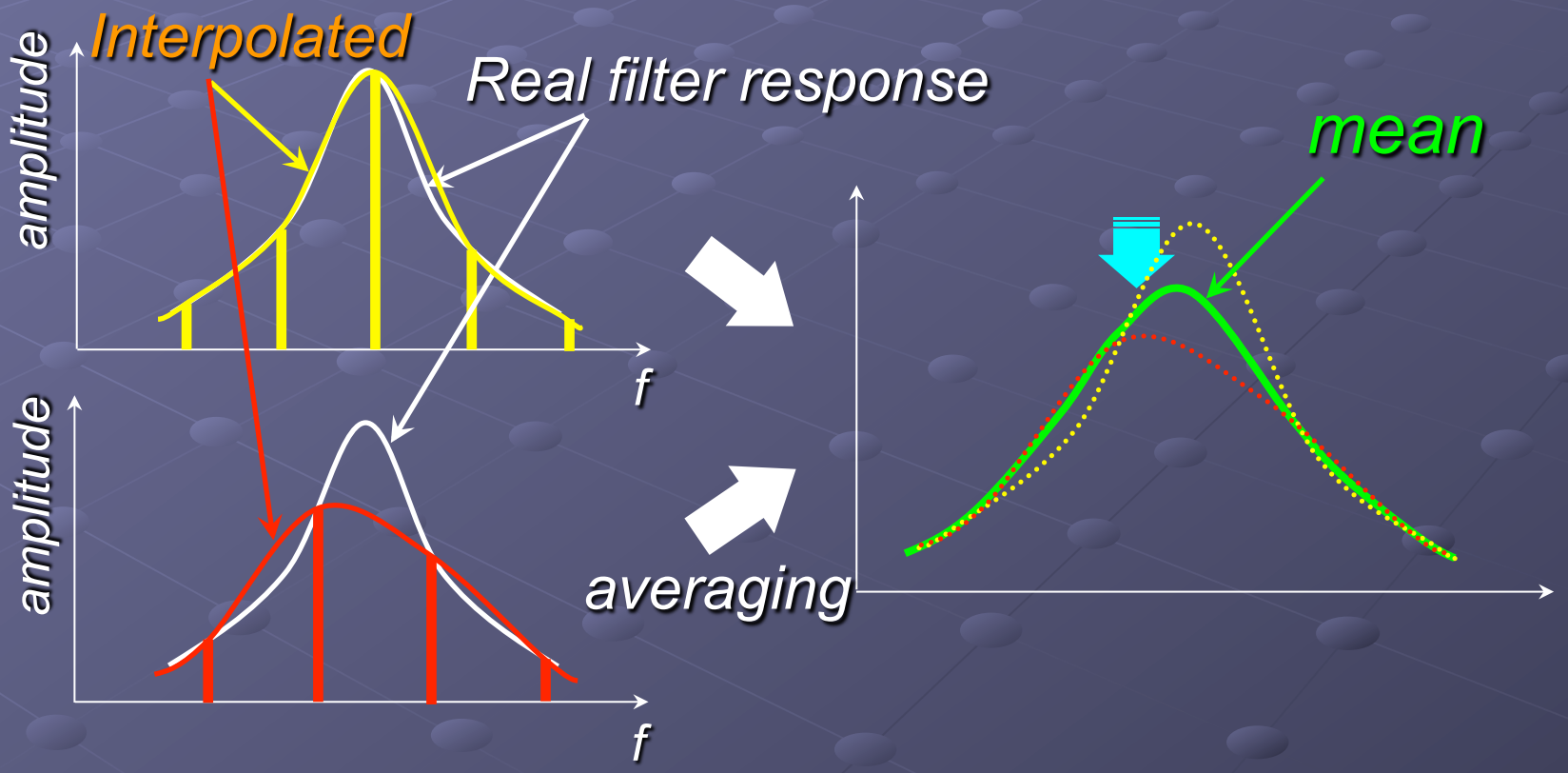


# Summary (2)

- The estimation using cepstral-domain criteria leads to larger distortions than the proposed method.
  - In parameterization, the criteria should be defined based on harmonics.

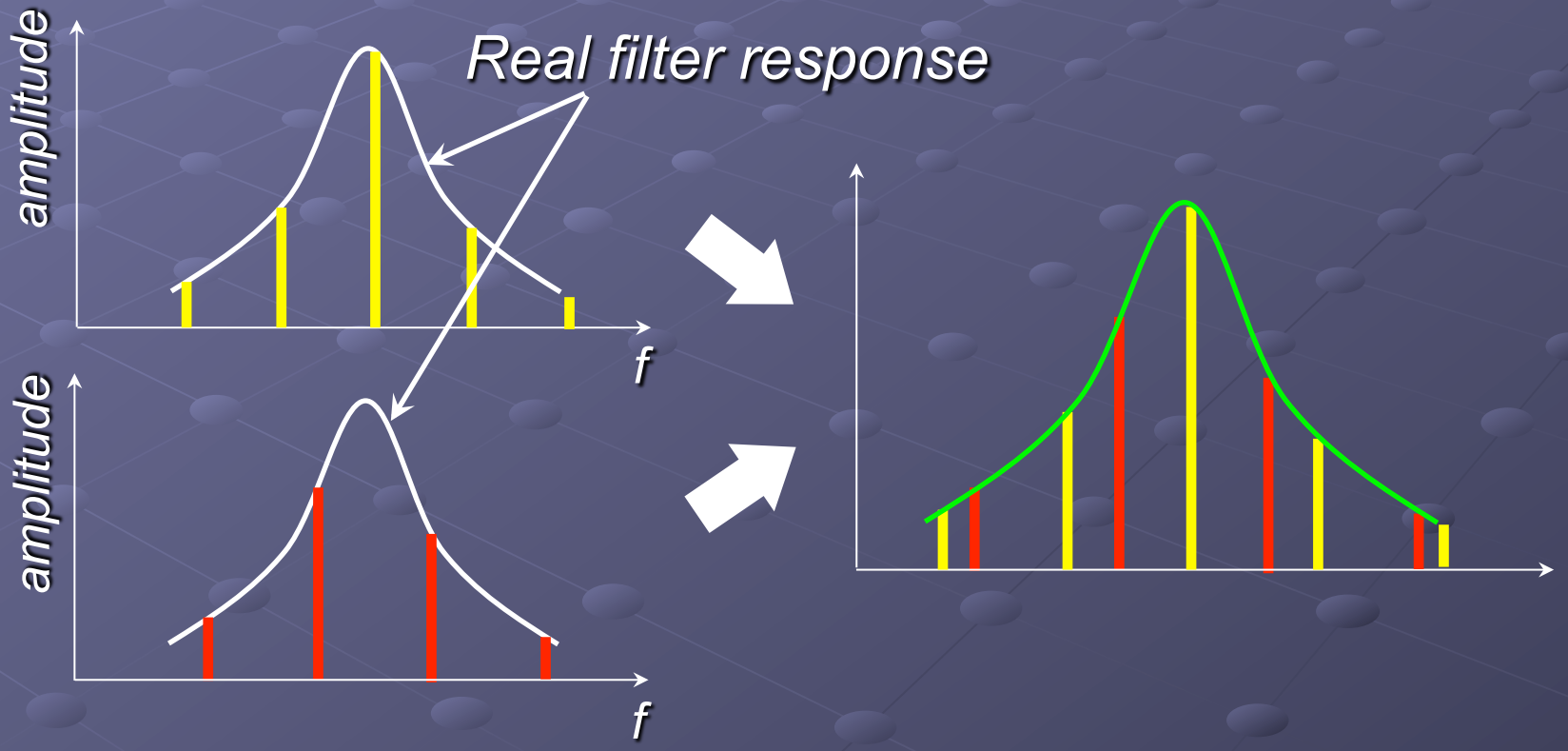
# Discussion

- Conventional methods treat reliable and unreliable characteristics equivalently.



# Discussion

- Proposed: Harmonics are first collected and then interpolated.



# Conclusion

- Conventional

*speech*

↓

parameterization

↓

*speech representation*
*(**interpolated** envelope)*

↕ *criterion*

$F_r(\mathbf{x})$

*estimated speech rep.*

- Proposed

*speech*

↓

harmonic analysis

↓

***harmonics***

↕ *criterion*

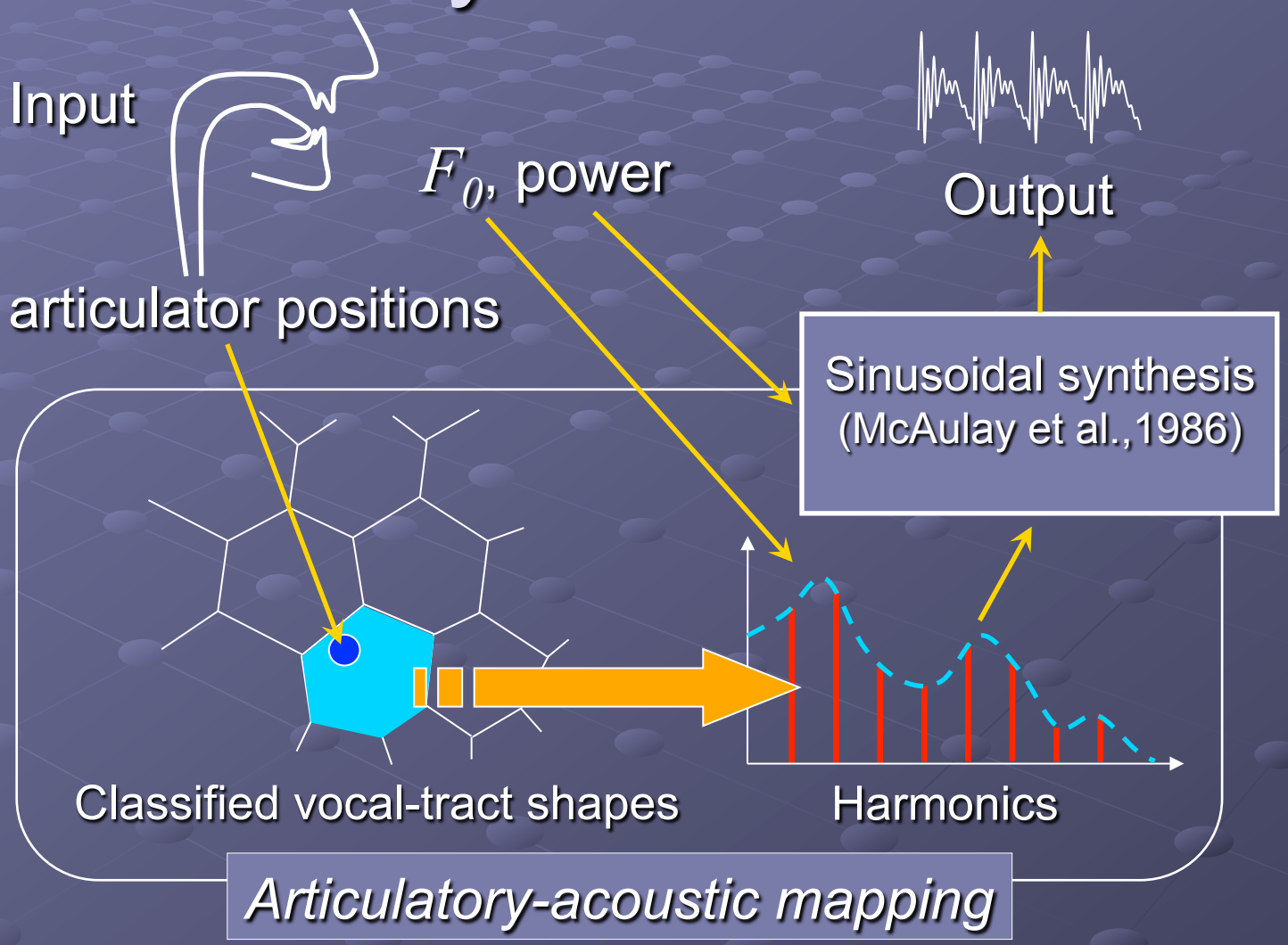
$F_h(\mathbf{x})$

*estimated **harmonics***

# Furthermore…

- Applying a source-filter separation (*Shiga et al*. 2003) further improves the estimation accuracy.
- A GMM-based mapping contributes toward improving the accuracy, and reducing acoustic discontinuity at cluster boundaries.

# Articulation-to-speech synthesis



# Articulation-to-speech synthesis demo

- “Are your grades higher or lower than Nancy's?”

| | PCM | PLM |
|---|---|---|
| conventional |  |  |
| proposed |  |  |
| prop.+ noise model |  |  |
| original |  | |

- “Stimulating discussions keep students' attention.”

| | PCM | PLM |
|---|---|---|
| conventional |  |  |
| proposed |  |  |
| prop.+ noise model |  |  |
| original |  | |

# Articulation-to-speech synthesis demo

- “The legislature met to judge the state of public education.”

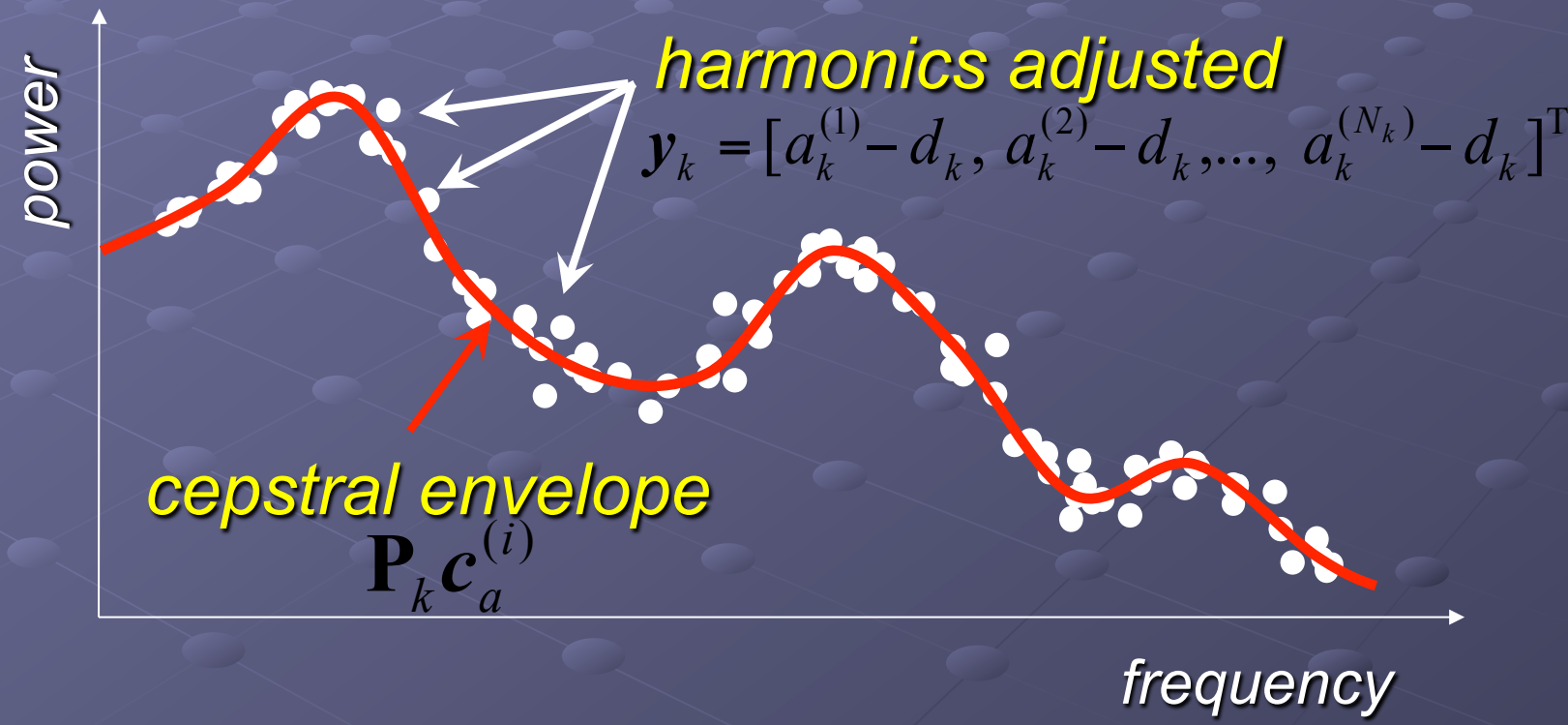
| | PCM | PLM |
|---|---|---|
| conventional |  |  |
| proposed |  |  |
| prop.+ noise model |  |  |
| original |  | |

- “Did you eat lunch yesterday?”

| | PCM | PLM |
|---|---|---|
| conventional |  |  |
| proposed |  |  |
| prop.+ noise model |  |  |
| original |  | |

# Piecewise constant mapping

- Finding cepstrum $\boldsymbol{c}_a^{(i)}$ representing power envelope for cluster $i$, $C^i$



# Piecewise constant mapping

- Finding cepstrum $\boldsymbol{c}_a^{(i)}$ representing power envelope for cluster $i$, $C^i$

$$E_a^{(i)} = \sum_{k \in C^i} \left[ \boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)} \right]^{\mathrm{T}} \mathbf{W}_k \left[ \boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)} \right]$$

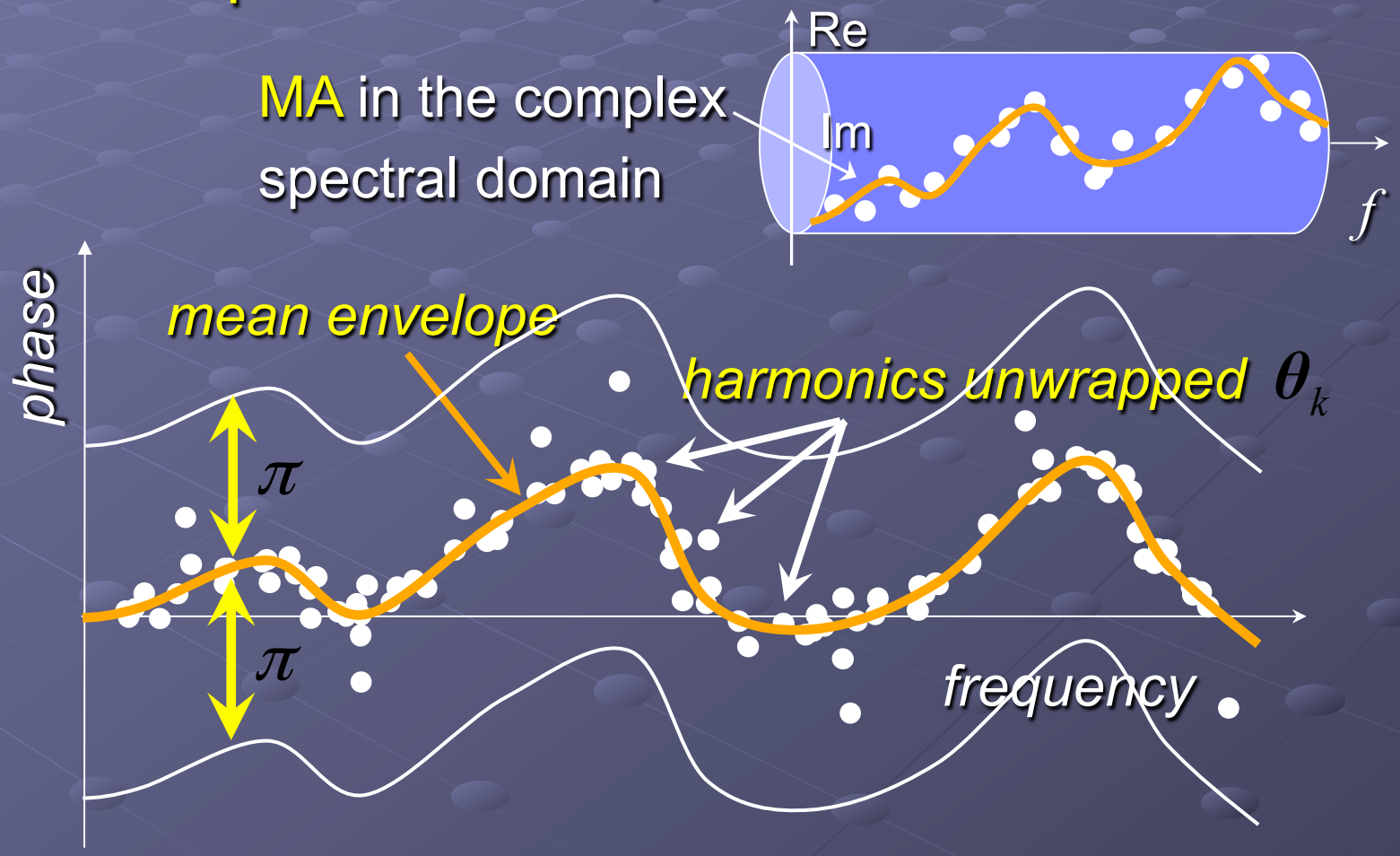
$$\boldsymbol{y}_k = [a_k^{(1)} - d_k,\ a_k^{(2)} - d_k, \ldots,\ a_k^{(N_k)} - d_k]^{\mathrm{T}}$$

$a_k^{(i)}$ : power of $l$-th harmonic in frame $k$

$$\mathbf{P}_k = \begin{bmatrix} 1 & 2\cos\Omega_k^{(1)} & 2\cos 2\Omega_k^{(1)} & \cdots & 2\cos p\Omega_k^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\Omega_k^{(N_k)} & 2\cos 2\Omega_k^{(N_k)} & \cdots & 2\cos p\Omega_k^{(N_k)} \end{bmatrix}$$

# Piecewise constant mapping

- Finding cepstrum $c_p^{(i)}$ representing phase envelope for cluster $i$, $C^i$



# Harmonic-based approach

- Finding cepstrum $\boldsymbol{c}_p^{(i)}$ representing phase envelope for cluster $i$, $C^i$

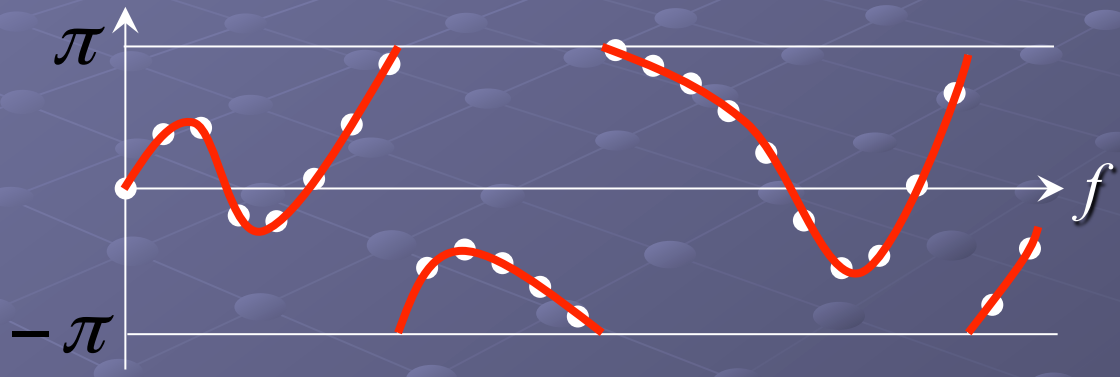
$$E_p^{(i)} = \sum_{k \in C^i} \left[\boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)}\right]^{\mathrm{T}} \mathbf{W}_k \left[\boldsymbol{\theta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)}\right]$$

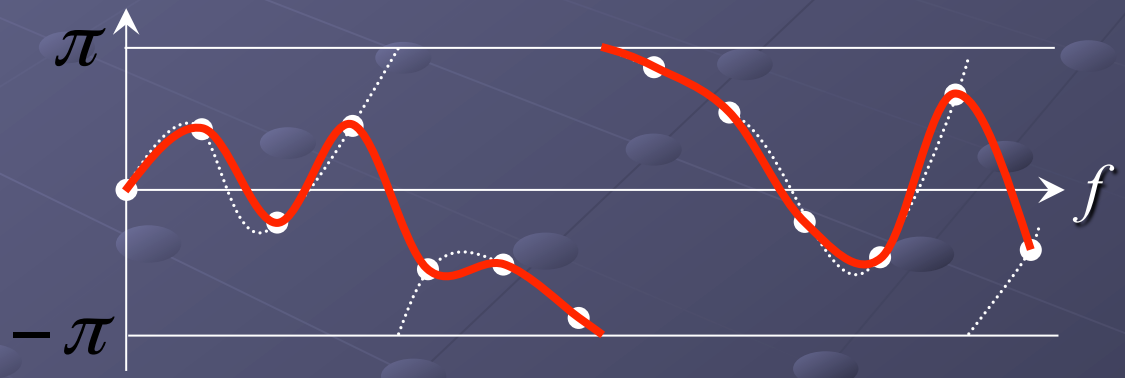
$$\boldsymbol{\theta}_k = (\text{unwrapped harmonic phase})$$

$$\mathbf{Q}_k = (-2) \cdot \begin{bmatrix} \sin \Omega_k^{(1)} & \sin 2\Omega_k^{(1)} & \cdots & \sin p\Omega_k^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \sin \Omega_k^{(N_k)} & \sin 2\Omega_k^{(N_k)} & \cdots & \sin p\Omega_k^{(N_k)} \end{bmatrix}$$

# Unwrapping problem of phase

- X. Huang, A. Acero and H.-W. Hon, *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*, Prentice Hall (2001)



# Comparison between a minimum phase spectrum and a spectrum estimated by PCM